CTS PROGRESS AT LIMSI

J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen

RT03 meeting
Boston, MA
May 19, 2003

TALK OUTLINE

- System overview
- Progress
- Acoustic models and decoding
- Language models and components for system combination
- Conclusions

SYSTEM OVERVIEW

RT02	RT03
Acoustic modeling	
PLP frontend	+ MFCC + PLP-S
Normalisations: VTLN, CMN, CVN	+ gender-dependent VTLN
MLE/MAP trained GD models	+ MMI training
28k tied-state triphones, 11k states	32k triphones
Cell models and cell switch	only one model set
Training data: SWB1, CallHome	+ CTRAN SWB2 data
48 phone symbols	+ reduced 35 phoneset

SYSTEM OVERVIEW

RT02	RT03
Language modeling	
42k vocabulary, ∼300 compounds 4-gram backoff LM Neural-net LM	50k vocabulary Improved LM (smoothing, data, ...)
Decoding	
3-gram lattices, 4g rescoring 3 passes decoding 2 and 5 phone class MLLR Consensus decoding with pron probs Confidence scores from CN	2-gram lattices, 4g rescoring 4 passes + 1 pass per component + 8 phone class MLLR

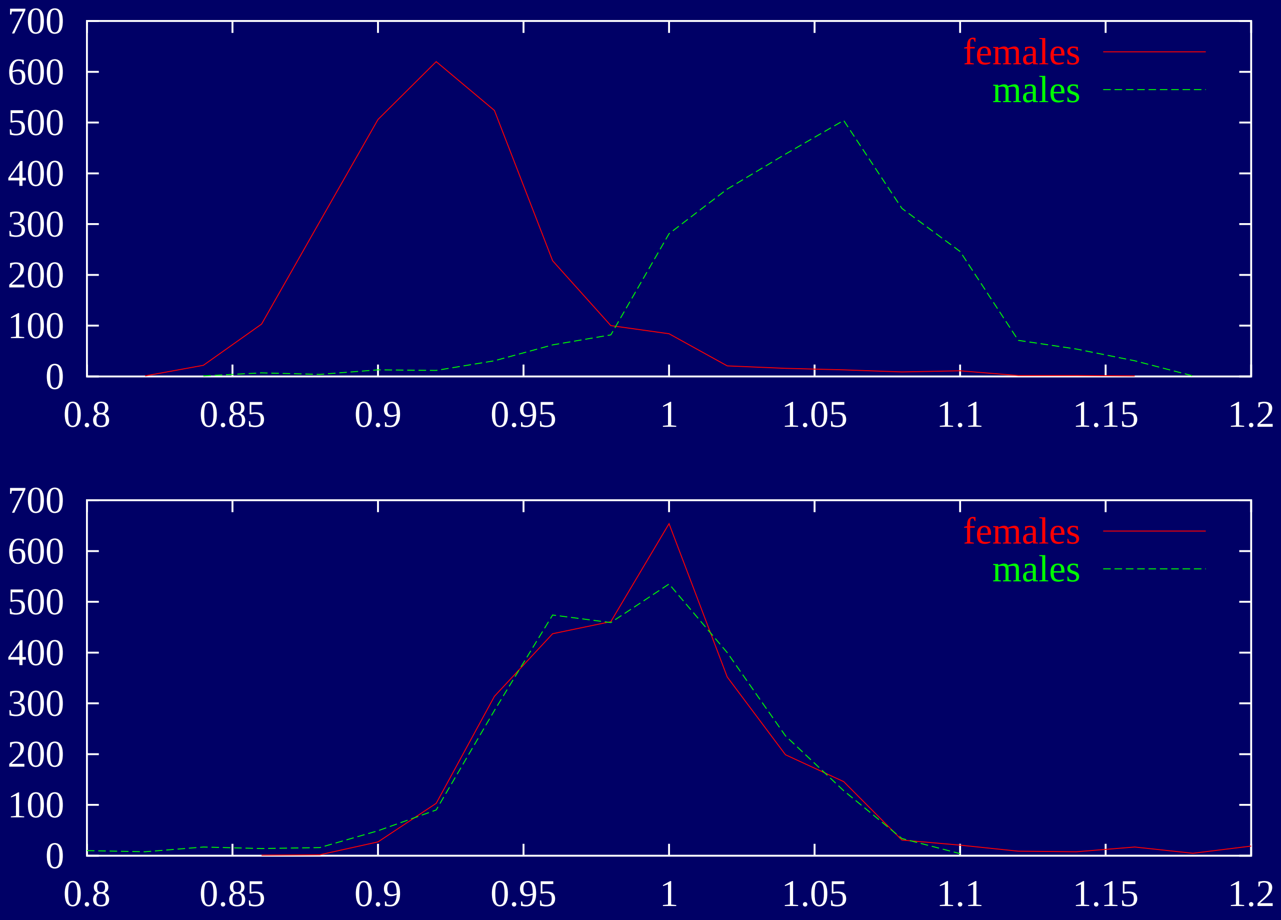
MAIN IMPROVEMENTS FOR RT03

- Gender-dependent VTLN (∼0.5%)
- MMI training of GD acoustic models (∼1.2%)
- Revised decoding (∼1.0%)
- CTRAN acoustic data (∼0.5%)
- Improved LM (∼1.0%)
- System combination (3 front-ends, 2 phone sets) (∼1.0%)
  Total gain ∼5.5% (eval01 and eval02)
- Integrated system with BBN

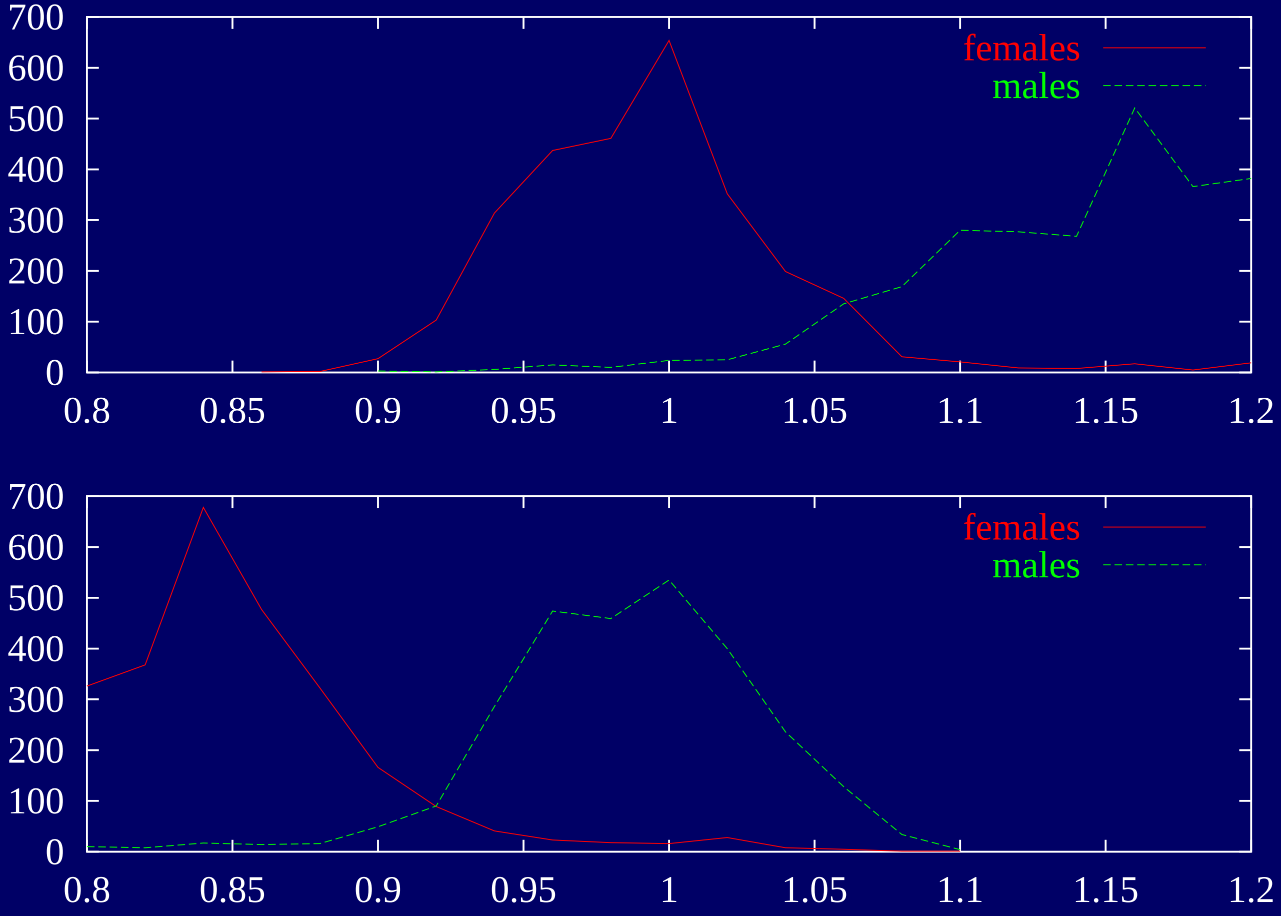
GENDER-DEPENDENT VTLN

- RT02 Method
  - Warp filter bank with a piecewise linear scaling function
  - GI ML estimation based on a 1st hypothesis with large models
  - Incremental search with 0.2 step
- RT03 Method
  - Gender-dependent warping (like 2 frontends)
  - ML estimation with single Gaussian models, Brent’s search
  - WER reduction $\sim$0.5%
  - Very fast (0.1xRT)

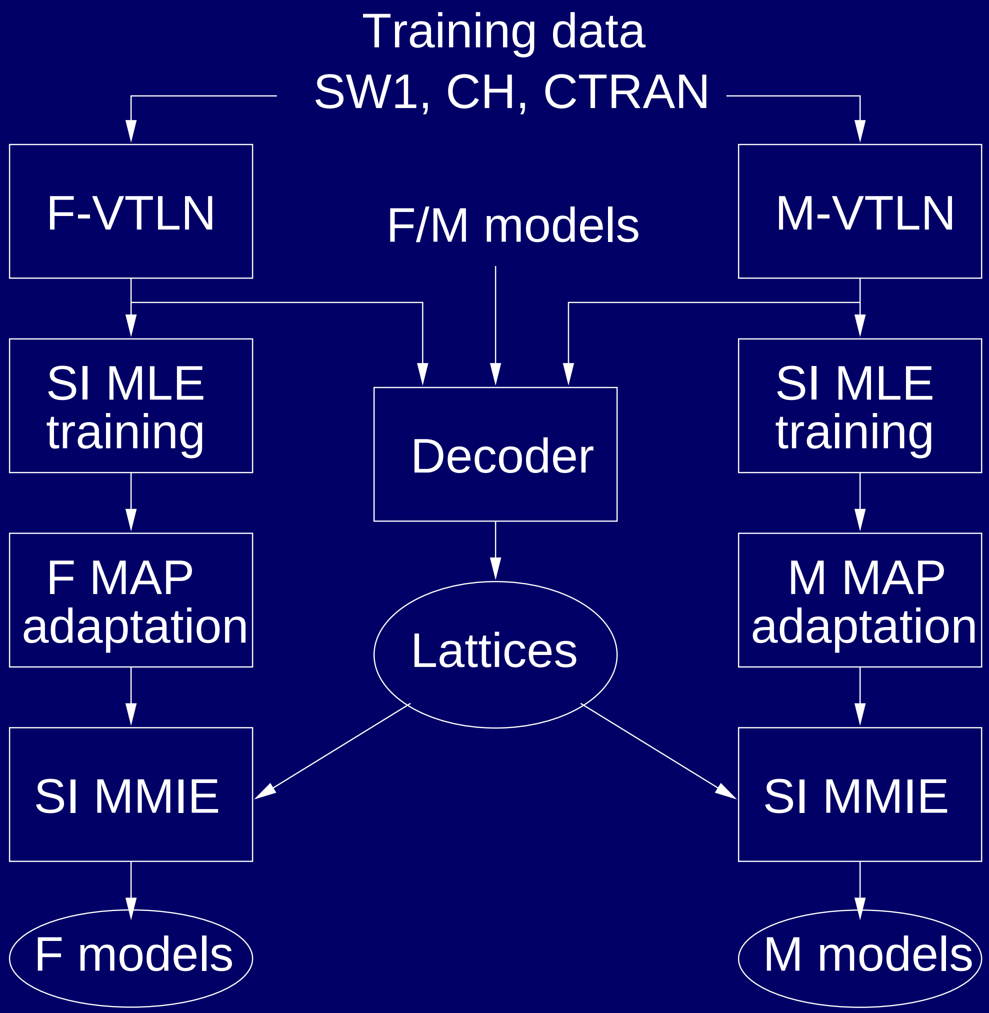
SI VTLN VERSUS GD VTLN



GD VTLN FOR MAP FM TRAINING



ACOUSTIC MODEL TRAINING



Training data
SW1, CH, CTRAN

F-VTLN

F/M models

M-VTLN

SI MLE training

Decoder

SI MLE training

F MAP adaptation

Lattices

M MAP adaptation

SI MMIE

SI MMIE

F models

M models

RT03 DECODING

	VTLN	*MLLR*	*LM*	*Eval01*	*Eval02*
Pass 1 PLP MLE	n	-	3g	35.6	40.5
Pass 2 PLP MMI	y	-	4g	25.2	29.0
Pass 3 PLP MMI	y	2	4g	22.8	26.2
Pass 4 PLP MMI	y	5	NN 4g	21.9	25.1
4 system combination	y	5/8	NN 4g	20.9	24.0

LANGUAGE MODEL AND COMPONENTS FOR SYSTEM COMBINATION

Presented by H. Schwenk

LM TRAINING CORPORA

RT02 system:

- SWB transcriptions from LDC (2.75M words) and from ISIP (2.93M words)
- CallHome corpus (229k words)
- SwitchBoard cellular transcriptions (217k words)
- BN commercial transcriptions (270M words)
- ”Switchboard-like” part of BN transcriptions (65M)

Additional corpora for RT03:

- CTRAN data from BBN [80h of fast SWB transcriptions] (1.1M words)
- WEB data from University of Washington (59M words)
- CNN television broadcast transcriptions [1/2000 - 3/2003] (80M words)

CONTRIBUTION OF NEW TEXTS

	Number of			*Perplexity*		
Language Model	*2-gram*	*3-gram*	*4-gram*	*Std*	*Decomp*	*WER*
RT03 dryrun	12M	21M	12M	82.8	60.2	22.94
+ CTRAN data	12M	21M	13M	80.8	58.8	22.76
+ improved smoothing	12M	22M	15M	80.3	58.4	22.47
+ WEB and CNN data	14M	24M	18M	79.3	57.8	22.21

Full decode with best acoustic models (without NN LM)

- Overall gain of 0.7%
- Important need for in-domain data

LANGUAGE MODELING FOR FISHER DATA

New Fisher data:

- Different epoch than previous CTS data
- New conversation topics
- No representative development data available

We tried to anticipate changes by updating the system vocabulary and the language models

- Added frequent words in recent BN data (mainly CNN)
- New wordlist: 51077 words (262 compounds), OOV 0.23% on eval01
  Eval03 LM has 16M 4-grams, 35M 3-grams and 22M 4-grams
- No change in px (55.6→55.3) and WER (21.92%→21.86%)

NEURAL NETWORK LANGUAGE MODEL

Characteristics:

- Performs n-gram probability estimation in a continuous space
- Trained only on the HUB5 corpora, interpolated with backoff LM
- Used for lattice rescoring during the last decoding pass

Performance comparison:

- Perplexity on eval01: 57.5 ⟶ 55.3 (68.8 ⟶ 63.5)

WER (%)	*Eval01*	*Eval02 (man)*	*Eval02 (auto)*	*Eval03 (auto)*
backoff LM	22.27	25.50	26.03	24.78
neural LM	21.86	25.09	25.71	24.43

Consistent gains of about 0.4%

COMPONENT SYSTEMS FOR COMBINATION

Four different systems were developed:

- **PLP** baseline system
- **PLP-S** short term cepstral mean and variance normalization.
- **PLP-R** reduced phone set (35 instead 45)
- **MFCC** front-end

Characteristics:

- All models are gender-dependent and MMI trained
- The alternate systems are built on top of the baseline system using 5-class MLLR adaptation

PERFORMANCE OF COMPONENT SYSTEMS

System	*Eval01*	*Eval02 (man)*	its Eval02 (auto)	*Eval03 (auto)*
PLP	21.9	25.1	25.7	24.4
PLP-S	21.8	25.0	25.6	24.3
MFCC	21.8	25.0	25.6	24.3
PLP-R	21.9	24.9	25.6	24.4

System combination with BBN:

- Analysis of several combinations of 8 systems (4 LIMSI, 4 BBN)
- Selected 2 LIMSI systems: **PLP-S** and **PLP-R**
- Rover followed by 8-class transatlantic MLLR adaptation of the individual systems
- Experimental details will be given in BBN’s talk

CONCLUSIONS

- Significant improvements compared to RT02
- Main changes: VTLN, MMI, front-ends, phone set, LM, decoding
- Total gain of about 4.5% on each component system
- Selected components for system integration with BBN
- Large gain with BBN+LIMSI integrated system (... coming soon in the BBN talk)